Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 96 (2013) 798 – 806

Procedia
Social and Behavioral Sciences

13th COTA International Conference of Transportation Professionals (CICTP 2013)

Correlation Analysis and Data Repair of Loop Data in Urban Expressway Based on Co-integration Theory

Liu Heng[a], Duan Zhengyu[a,b,*], Shi Xiaofa[a]

[a]*School of Transportation Engineering, Tongji University, Address, 4800 Cao'an Road, Shanghai, 201804, China*
[b] *Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University,Address, 4800 Cao'an Road, Shanghai, 201804, China*

Abstract

Correlation between road sections has a strong practical significance in traffic raw data repair, precision control and traffic prediction. In this paper, the co-integration theory of econometrics is applied to correlation analysis of the traffic flow parameters series based on loop detection data of Shanghai expressway. Then we establish data repair model according to the plane geometry types of road, based on the correlation between road sections. Experimental instances show that compared with traditional approaches, this method can reduce the error of data repair, and has a good practical prospect.

Keywords: co-integration theory; loop detection data; correlation analysis; data repair ;

1. Introduction

There is a temporal and spatial association between traffic characteristics of the urban expressway sections, due to the impact of factors such as the characteristic of the inhabitant trip OD, road network structure. Spatial-temporal correlation analysis of urban expressway can help us understand evolution of road traffic status, identify the different characteristics of road traffic conditions in different regional, and grasp the overall operation status and clustering features of the urban road network more accurately. In addition, spatial-temporal correlation analysis of road cross-sections provides a good foundation for detector data repair, traffic prediction, and traffic control.

In recent years, researchers paid more attention on correlation analysis in traffic science, and accumulated some results and experiences. Based on the data of travel speed, Xiang He studied the temporal-spatial correlation

* Corresponding author. Tel.: 86-21-69583775; fax: 86-21-69583775.
E-mail address: d_zy@163.com

doi:10.1016/j.sbspro.2013.08.091

of road section in the normal traffic situation, through the method of Random Matrix Theory and spatial statistics. Furthermore, he discussed that the changes of the clustering phenomena along with time, which could be used in analyzing the characteristic of vehicle trip in cities at different time. [1] Linlin Liang stated that the temporal-spatial banded distribution of traffic congestion is apparent and the correlation of expressway congestion is slightly stronger than that of the independence. His study is based on loop detection data and congestion association rules. [2]

Nowadays, with the developments of information technology and traffic information collection system, different kinds of continuous traffic data are available to transportation researchers. Studies on this subject have attempted to get high quality repairing-data and hunt for numerous extensions of the methodology.[3][4][5] Ya Sun studied the methods to improve data quality of fixed collection systems. In his study, fixed collection data quality control theory and methods were discussed and the missing data, irregular time data, the erroneous data distinction model and the revision model were established.[6]

Co-integration theory has been used for econometrics research, [7][8] while rarely been applied in traffic engineering. In Weifeng Li and Zhengyu Duan's study, the co-integration theory and error correction model of econometrics are applied to the data fusion method to construct error correction model of floating car data.[9] In this paper, the co-integration theory of econometrics is applied to correlation analysis of the traffic flow parameters series, and the conclusion of the correlation between road cross-sections is applied to missing traffic flow data repair.

2. Integration and Co-integration Theory

Most time-series variables are not stationary. The variation of a nonstationary series are different and random at all-time points, which results in the variation of its numerical characteristics with time, such as mean and standard deviation. Thus the methods which could be applied to stationary time-series data are inapplicable. Test for a unit root is a simple and rapid way to judge the stationarity of a time-series. Common methods include Augmented Dickey-Fuller test (ADF test), Dickey-Fuller test with GLS (DFGLS test), Phillips-Perron test (PP test), etc.

2.1. Integration

Some nonstationary time-series, which is called integration, can reduce them to stationarity by use differencing algorithm. The process that translates nonstationary series into stationary series by differencing is defined as difference-stationary process.

If y_t is a nonstationary series, a is a constant, and u_t is a white noise series, then first difference of series y_t is stationary.

$$\Delta y_t = y_t - y_{t-1} = a + u_t \ . \tag{1}$$

The series y_t is said to be integrated of order one, denoted $y_t \sim I(1)$, because taking a first difference produces a stationary process. A nonstationary series is integrated of order d, denoted $I(d)$, if it becomes stationary after being difference d times.[10]

2.2. Co-integration Theory

If linear combination of two or more nonstationary series is stationary, then this linear combination will be denoted by co-integrated equation, which can be used to depict the long-term stable equilibrium relationship of two or more series.

The components of a k dimension vector $Y(y_1,y_2,\cdots y_t)'$ are said to be integrated of order d,b, denoted $Y\sim CI(d,b)$, if the vector Y satisfies the following two requirements:

(1) $Y\sim I(d)$, which means that every component of Y is amenable to be integrated of order d, that is, $y_t\sim I(d)$;

(2) non-zero vector β exists, which makes $\beta' Y\sim I(d-b)$, and $0<b<=d$.

The vector Y is said to be co-integrated for short, and the vector β is a co-integrating vector.[11]

Take a fully specified regression model (2) as an example. If y_t and x_t are $I(1)$, there may be a β such that ε_t is $I(0)$. The implication would be that the series are drifting together at roughly the same rate. y_t and x_t are said to be co-integrated.

$$\varepsilon_t = y_t - \beta x_t . \tag{2}$$

In addition, if two or more series are integrated to different order, then the linear combinations of them will be integrated to the higher (or the highest) of the orders.

The purpose of the study co-integration can be divided into two: one point is to judge whether there is a co-integration relationship between a group of non-stationary series, another point is to determine the rationality of the design of the linear regression equation by the co-integration test. The main idea and process of these two points are exactly the same. It's important to note that the non-stationary series must be integrations with same order.

3. Data description and analysis

3.1. Data description

The east part of North-south Viaduct in Shanghai viaduct is selected, and the study region is from Ruban Road Overpass to Gonghexin Road Overpass, stretching nearly 7.5 kilometers long and including 20 sections. In view of the quality and completeness of the data and the representativeness of the study, the paper conducts an empirical study of the correlation between road sections and data repairing by taking March 24, 2010(Wednesday) all day long's loop detection data as example. Loop vehicle detectors record a data every 20 seconds, includes the traffic flow, average speed of vehicles pass through the detector, detector's occupancy, etc. According to some existing research on roads' correlation, it's said that traffic flow has more apparent dominance than speed data at research on roads' correlation. So we analyze the correlation of the 20 sections based on traffic flow series.

As we can see in figure 1, the relationship between section n and other sections is the object of study in this paper. And *section(+i)* is defined as the section n+1, which means the downstream section. For example, *section(-1)* means *section 5* and *section(+2)* means *section 8* when we are analyzing the correlation between *section 6* and other sections.

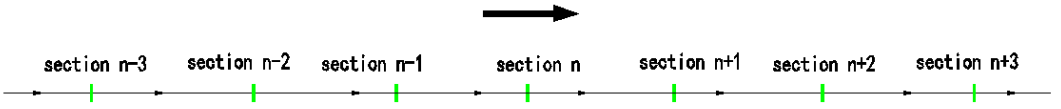


Fig. 1. Sections' definition

3.2. Series' stationarity analysis

We study every section's accumulative traffic flow respectively in accordance with 1min, 2min, 5min and 10 min. Analyzing correlation coefficient of the four accumulative flow series, the conclusion is in Table 1. $\rho(x_n x_{n+i})$ represents the correlation coefficient of flow series between section n and section $n+i$, and $n=1,2\cdots20$, $i>0$. What's more, ρ_a means the sum $\rho(x_n x_{n+i})$ of the 20 sections.

Table 1. sections' correlation coefficient

$\rho(x_n x_{n+i})$	±1	±2	±3	±4	ρ_a
1 min	0.9450	0.9184	0.9031	0.8870	0.8584
2 min	0.9731	0.9530	0.9371	0.9200	0.8863
5 min	0.9863	0.9725	0.9575	0.9400	0.9021
10min	0.9898	0.9793	0.9665	0.9504	0.9111

As we can see in Table 1, all of the correlation coefficients are larger than 0.85, which means that the flow of the sections is significantly correlative. And the correlation coefficients of two sections' flow series reduce when distance increases (here, distance shown as the value of i.), which shows that their correlation coefficients are depressed for the reason of speed, on-ramp and down- ramp, etc. In addition, the correlation coefficients of 1 min, 2min, 5 min, 10 min significantly increase in turn, and Correlation enhance, which shows that the relationship of sections' traffic flow series changes from short-term imbalance(random disturbance) to long-term equilibrium.

Then, 2 min is selected as analysis interval based on a comprehensive consideration of short-term imbalance and long-term equilibrium. We find that the graphics of traffic flow series' first-order difference are similar to white noise series, so we determine the first-order difference series initially as stationary series. Take the 6th section as an example, Fig 2 and Fig 3 respectively show the 6th section's traffic flow series x_6 and its first-order difference series $d(x_6)$.

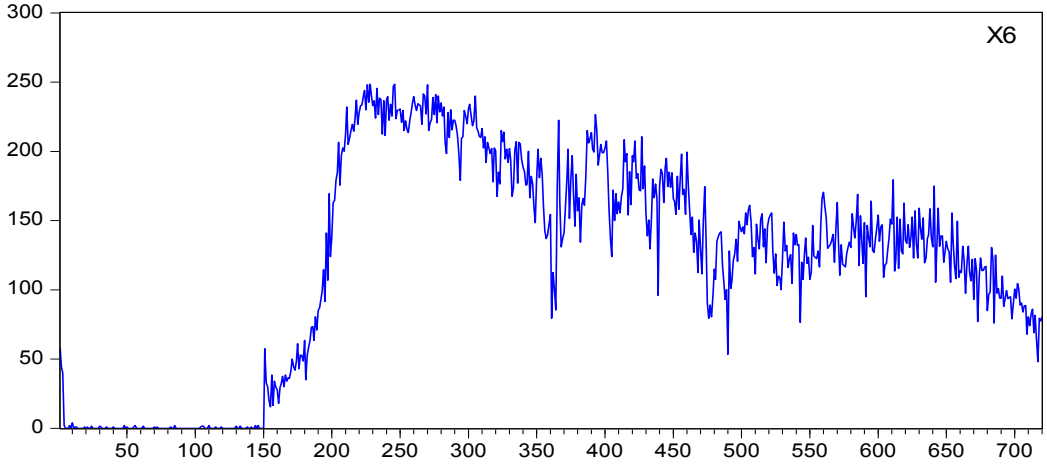


Fig. 2. the 6th section's traffic flow series x_6

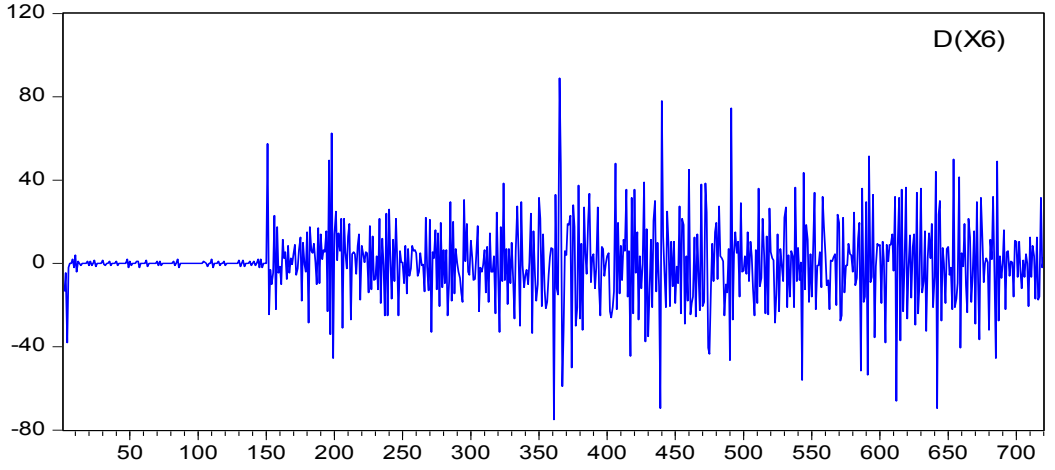


Fig. 3. the 6th section's first-order difference traffic flow series $d(x_6)$

20 sections' flow are non-stationary series, and they are integrated of order one (x_n~$I(1)$,n=1,2,⋯20) by ADF test. 2-minute cumulative flow of the 6th section, for instance, as shown in the Table 2.

Table 2. the 6th section's traffic flow series stationarity analysis

series	t-Statistic	Test critical values(1% level)	Test critical values(10% level)	Prob.*	lag length	stationarity
x_6	-1.074331	-3.970976	-3.130355	0.9312	5	no
d（x_6）	-17.28751	-3.439268	-2.568864	0	4	yes

3.3. Series' co-integration analysis

The Engle and Granger method is used for testing for the traffic flow series' co-integration, and it's based on assessing whether single-equation estimates of the equilibrium errors appear to be stationary. There are two steps for the test: step one is to establish the equilibrium relationship that is, co-integration function, and the least square method is usually used. If the function is correct and OLS estimator is consistent, and then is the step two, estimating stationarity and integration of residuals from this regression from this regression function. For the traffic flow series are integrated to order one, we can distinguish between co-integration relationships between different sections' traffic flow.

The paper tests for the co-integration relationship between a section and its adjacent three sections, the co-integration function can be written

$$x_{n+i}=\beta_n x_n+u_n \tag{3}$$

here $n=1,2,\cdots 20$, and $i=1,2,3$. The test results are shown in Table 3.

Table 3. regression parameters of 2-minute cumulative flow series' co-integration

	β_n	Prob.*	u_n	Prob.*	Adjusted-R^2	stationarity
i=1	0.979749	0.0000	4.465886	0.1187	0.9470	yes

i=2	0.964236	0.0000	8.546005	0.0431	0.9083	yes
i=3	0.992056	0.0140	12.73657	0.0695	0.8785	yes

4. Correlation Analysis of road sections and data repairing

4.1. Correlation Analysis of road sections

According to co-integration analysis of the road sections, section *n* and section(±1)，section(±2)，section(±3) appear to be significantly co-integrated and correlative. Compared with *i=2* and *i=3*, the model (3) showed a better model-fitting degree when *i=1*. Putting this and the correlation coefficient analyzed before, we can come to the following conclusions: successive sections vary with less random shocks and appear to be much more significantly co-integrated, and the correlation between them is much larger. Therefore, the following parts will only analyze the relationship between them.

Analyzed the co-integration of the successive sections, the result is as following:

(1) $\beta_n>0$, which means there is a positive correlation between the successive sections;

(2) the mean $\beta_n\approx1$, which means their rate of change are almost on a par, and the traffic flow of the North-south Viaduct is stationary;

(3) the larger β_n is, the more sensitive the successive sections are. $\beta_9=1.779083$ is the maximum of β_n and $\beta_{18}=0.485085$ is the minimum, which means the sensibility index of section 9 and section 10 is the highest, and that of section 18 and section 19 is the least.

4.2. Sections' data repairing

According to the correlation analysis of the sections, in the paper we select road section rather than a single loop vehicle detector as the research unit, and put forward two kinds of different traffic flow data repairing model based on the co-integration function.

(1) Successive sections' data is lost or corrupted

If some or all detectors of two successive sections are broken, resulting in data loss or corruption, then co-integration function could be used for data repairing directly. Combined with the proceeding results in Table 3 and correlation analysis of road sections, successive sections' co-integration function is applied to data repairing, which would increase the accuracy. The model is as following,

$$x_{n\pm1}=\beta_n x_n+u_n. \tag{4}$$

Here we repair all the sections' data based on their upstream section except *section 1*. Results are shown in Table 4. Compared with regression analysis based on the adjacent lane, this model is more practical.

Table 4. accuracy and error of data repairing model (4)

section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）	section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）
1	6.859368	9.624214	11	16.32004	47.20898
2	7.103647	12.85841	12	10.33479	21.12529
3	10.11567	24.87304	13	9.475411	14.37492
4	7.652687	9.644003	14	26.82849	33.81344
5	11.66718	38.11614	15	11.50549	13.27162
6	10.29061	11.96552	16	13.79785	15.68284

7	8.601930	10.62213	17	8.194586	6.808930
8	8.019206	15.44066	18	7.755582	6.374142
9	6.411589	12.61771	19	7.169776	12.13145
10	17.27083	52.70619	20	5.980993	13.31652

As shown in Table 4, some sections' Mean Absolute Percent Error is large. We try to repair these data based on their downstream section. Results are shown in Table 5.

Table 5. accuracy and error of data repairing model (4)-based on the downstream section

section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）	section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）
3	7.664657	13.57651	11	11.50549	13.27162
5	8.211905	20.28852	12	9.510015	13.56847
10	14.67477	13.69059	14	11.50549	13.27162

Compared with the results in Table 4, the accuracy is much higher in Table 5, which is because these 6 sections' traffic 2-min flow series are more correlative with their upstream section.

(2) Section's data is lost or corrupted, but both ahead section and behind one are available

As shown in Table 1, there are significant co-integration relation between section n and *section(±1)*, *section(±2)*, *section(±3)*. Refer to current research achievements at home and abroad, consider utilizing regression estimation based on some sections ahead and behind. The model is written

$$x_n = \sum_{i=1}^{k} (\beta_{n-i} x_{n-i} + \beta_{n+i} x_{n+i}) + c_n \tag{5}$$

In the function, x_n is estimation value of *section n*'s traffic flow, and $x_{n\pm i}$ is the observed value of *section(±i)*'s traffic flow. According to analysis of model-fitting degree, we can obtain a data repairing model with a high degree of accuracy and model-fitting degree when k=1 and c_n=0.

Table 5. accuracy and error of data repairing model(5)

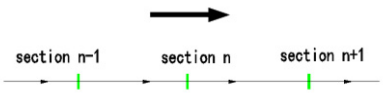
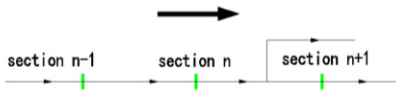
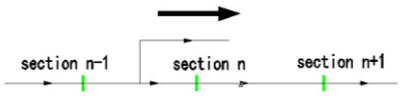
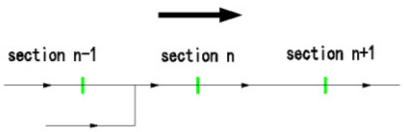
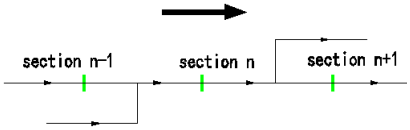
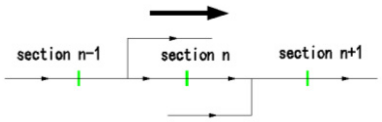
section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）	section	Mean Absolute Error(*pcu*)	Mean Absolute Percent Error（%）
2	6.452361	10.6423	11	8.773511	8.100038
3	7.341921	13.94321	12	7.109391	7.278932
4	6.644713	8.923212	13	9.060351	13.8867
5	7.258469	10.05757	14	13.65864	23.98289
6	7.505765	8.119332	15	10.66546	49.31832
7	6.833026	8.623771	16	7.560233	7.332849
8	5.32846	10.35473	17	5.205942	4.037739
9	5.333557	10.21608	18	6.804049	5.879158
10	10.75014	11.93596	19	5.931049	10.27502

As Shown in Table 6, the accuracy of all the 20 sections is quiet high, except for *section 14* and *section 15*. The average value of the 18 sections' Mean Absolute Percent Error is about 10%. Only *section 14* and *section 15*'s Mean Absolute Percent Error are over 20%. We analyze *section 14* and *section 15* further by looking at their geographic location and the plane geometry types of road nearby, and find that they are respectively at the both sides of the Tianmu Road Overpass. On-ramps and down- ramps are between section 13 and section 14, and

between section 14 and section 15, but none between section 14 and 15. What's more, there are large numbers of vehicles driving in and out the North-south Viaduct by Tianmu Road. Therefore, the correlation between *section 14* and *15* is much larger. But compared *section 13* with *section 14*, and compared *section 15* with *section 16,* volatility and uncertainty are higher. When repair one of these two sections, we will use the first data repairing model only based on the other section. As we can see in Table 4 and Table 5, when we repair *section 15*'s traffic flow data based on model (4), we can reduce the Mean Absolute Percent Error from nearly 49.32% to 13.27%. Above all, errors are much less than 17%~20%, which Ya Sun stated in his article

By statistical analysis, we find that β_{n-1}、β_{n+1}'s amplitude of variation is small, almost with the range from 0 to 1. Classify the successive three sections by the road structure and analyze their implied disciplinary change as Table 7.

Table 7. analyzing the value of β_n

plane geometry types of road	β_{n-1}: β_{n+1} (β_{n+1}: β_{n-1})	recommended value
	0.8~1.2	0.45:0.55 0.55:0.45
 	0.4~0.8	0.75:0.45
 	0.35~0.45	0.26:0.6
 	-	-

5. Conclusion

Reference to the co-integration theory of the non-classical econometrics, in the paper the correlation between road sections is analyzed, and the conclusion of correlation analysis is applied to data repair. The main conclusions are as follows:

(1) Taking the North-south Elevated Road in Shanghai as an example, we analyzes traffic flow data, indicates that road sections' traffic flow series meet the characters of co-integrated time series and there is a significant or greatly significant linear relationship between the successive sections.

(2) According to the co-integration function and correlation analysis, two kinds of data repairing models are built for different situations. And compared with traffic flow's repairing Mean Absolute Percent Error over 17%, the Mean Absolute Percent Error in the second model is only about 10%, it's proven that produce much more accurate results when repair traffic flow data. The models are practical for their high precision and easy operation.

(3) Classified the sections by their plane geometry types, the article studies the regularity that the range of coefficient β_n's value appears obviously, and recommends optimal value of coefficient β_n on the basis of different road structures, which guarantee the models keep high accuracy.

Acknowledge

This work is supported by the Open Project of Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University: Feature Extraction and Evolution Analysis Method for Traffic State of Regional Road Network.

Reference

[1] Xiang He. Temporal-spatial Correlation Analysis of Urban Road in Information Environment[D]. Shanghai: Tongji University, *2009.3*(5~10)

[2] Linlin Liang. Congestion Correlation Analysis of Urban Expressway Based on Data Mining[D]. Shanghai: Tongji University, *2012.3*(33~50)

[3] M Jain, B Coifman (2005), Improved Speed Estimates from Freeway Traffic Detectors, *Journal of Transportation Engineering*

[4] JIANG Gui-yan, GANG Long-hui, ZHANG Xiao-dong, WANG Jiang-feng (2004). Malfunction identifying and modifying of dynamic traffic data. *Journal of Traffic and Transportation Engineering*, 1. 3 - 5

[5] Chen.M., J. Xia, R. Liu. Developing a Strategy for Imputing Missing Volume Data, *Proceedings of the 85th Annual Conference of Transportation Research Board*, *2006.1*.

[6] Ya Sun. Study on Fixed Traffic Information Collection System Based on Data Quality [D]. Shanghai: Tongji University, *2008.7*(84~95)

[7] Ji Yushan, Wu Yongmin. On the Cointegration-model of the Relations between Industrial Structure and Economic Growth in China. *Contemporary Economic Research. 2006.6*

[8] Ya-qun HE, Guo-hong LAO, Chris E OSUCH, Wei-ran ZUO, Bao-feng WEN. Co-integration-based analysis of energy assurance for steady economic growth in China. J*ournal of China University of Mining and Technology. 2008.6* (pp. 250-254)

[9] Li Weifeng, Duan Zhengyu. Error Correction Model of Floating Car Data Based on Co-integration Theory.

[10] Tiemei Gao (2009). *Econometric Analysis and Modeling* (2th ed). Beijing: Tsinghua University Press (pp. 164-182)

[11] William H. Greene (2001).*Econometric Analysis* (4th ed). New York: Prentice Hall (pp. 748 - 760)